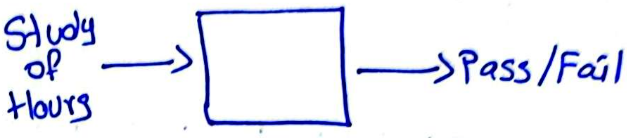
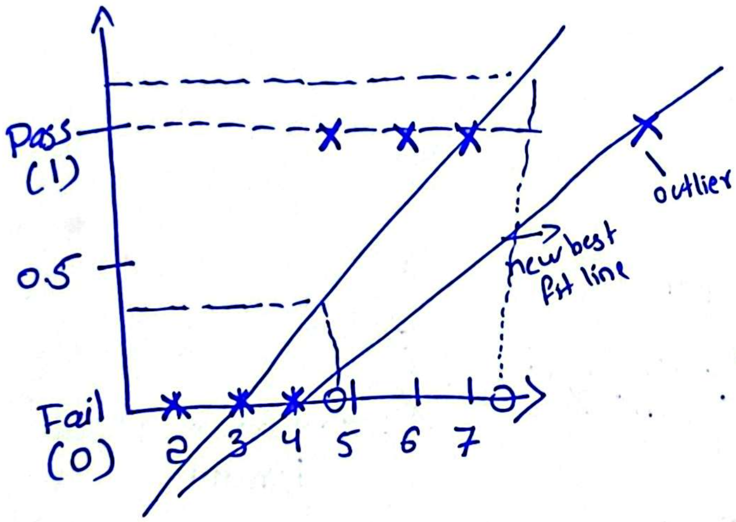
Logistic Regression

Dataset

Study Hours	Pass/Fail
2	Fail
3	Fail
4	Fail
5	Pass
6	Pass
7	Pass
12	Pass

→ outlier



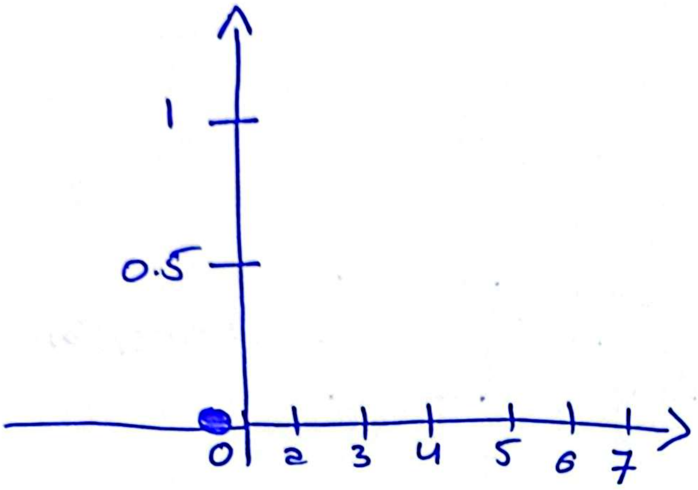


with Linear Regression

$$\begin{cases} \leq 0.5 \Rightarrow 0 \\ > 0.5 \Rightarrow 1 \end{cases}$$

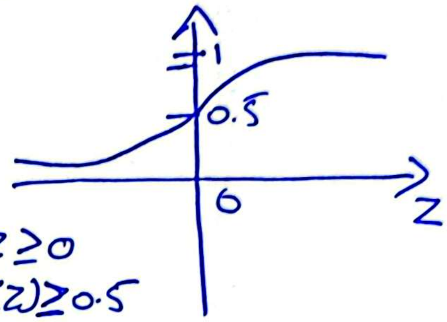
1) outlier → best fit line changes

2) >1 and <0 {squash line}



Sigmoid Activation Function

$$\frac{1}{1+e^{-z}}$$



$z \geq 0$

$\sigma(z) \geq 0.5$

$$h_\theta(x) = \sigma(\theta_0 + \theta_1 x)$$
$$= \sigma(z)$$
$$= \frac{1}{1+e^{-z}}$$

Linear Regression Cost Function

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x)^i - y^i\right)^2 \Rightarrow \text{Convex Function}$$

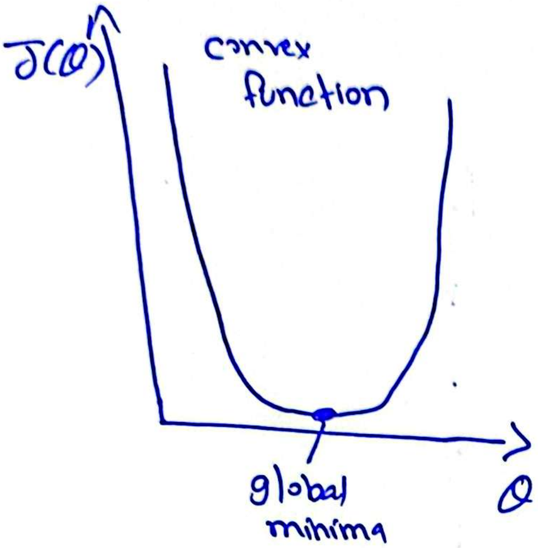
$$h_\theta(x) = \theta_0 + \theta_1 x$$

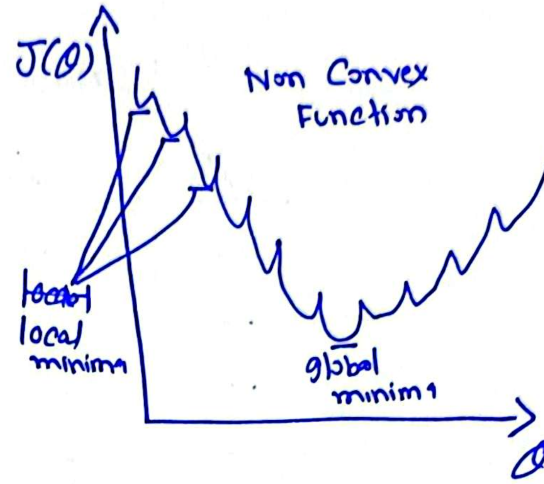
Logistic Regression Cost Function

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x)^i - y^i\right)^2 \Rightarrow \text{non convex function}$$

$\cancel{h_\theta(x) = \theta_0 + \theta_1 x}$ $\quad h_\theta(x) = \dfrac{1}{1+e^{-z}}$

$$z = \theta_0 + \theta_1 x$$





Logistic Regression Cost Function

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x)^i - y^i\right)^2 \qquad h_\theta(x) = \frac{1}{1+e^{-z}}$$

$$\Downarrow$$

$$\text{Cost}\left(h_\theta(x)^i, y^i\right) \qquad z = \cancel{h_\theta}\, \theta_0 + \theta_1 x$$

$$\text{Cost}\left(h_\theta(x)^i, y^i\right) = \begin{cases} -\log\left(h_\theta(x)\right) & \text{if } y = 1 \\ -\log\left(1 - h_\theta(x)\right) & \text{if } y = 0 \end{cases}$$

$$\text{Cost}\left(h_\theta(x)^i, y^i\right) = -y \log\left(h_\theta(x)\right) - (1-y) \log\left(1 - h_\theta(x)\right)$$

$$\underbrace{J(\theta_0, \theta_1) = -\frac{1}{2m} \sum_{i=1}^{m} \left[y^i \log\left(h_\theta(x)^i\right) - (1-y^i) \log\left(1 - h_\theta(x)^i\right) \right]}_{\text{Convex Function}}$$

Pe

→ Performance Metrics

Dataset

x_1	x_2	y	$\hat{y}$
–	–	0	1
–	–	1	1
–	–	0	0
–	–	1	1
–	–	1	1
–	–	0	1
–	–	1	0

> Confusion Matrix

Predicted values		1	0
	1	3	2
	0	1	1

⇒

	Actual value 1	0
1	TP	FP
0	FN	~~FN~~ TN

> Accuracy

$$\frac{TP+TN}{TP+TN+FP+FN} = \frac{3+1}{3+1+1+1} = \frac{4}{7}$$

> Precision and Recall

Dataset → 900 : 1, → 100 : 0 } Imbalanced Dataset

$$\text{Precision} = \frac{TP}{TP+FP}$$ } out of all the actual value how many are correctly predicted

$$\text{Recall} = \frac{TP}{TP+FN}$$ } out of all the predicted values how many are correctly predicted.

Actual   Spam Classification

		1	0
Predicted	1	TP	FP✓
	0	FN	TN

⇓

Precision: reduce FP

Mail → Spam
Model → Spam } good

Mail → Not Spam
Model → Spam } Issue

	1	0
1	TP	FP
0	✓FN	TN

⇓

Recall: reduce FN

Truth → diabetes
Model → no diabetes } Issue

> Fβ Score $(1+\beta^2)\dfrac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

$$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

if FP and FN are both equal important $\Rightarrow \beta = 1$

② if FP is more Important than FN $\Rightarrow \beta = 0.25$

$$F\beta \text{ Score} = 1 + (0.25) \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

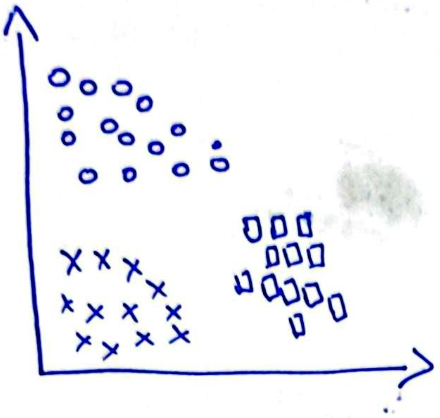
F(0.25)

③ if FN is more Important than FP

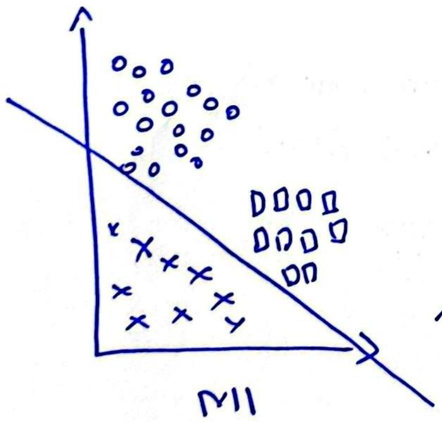
$\beta = 2$

$$F2 \text{ Score} = (1+2) \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

→ Logistic Regression (One Versus Rest)



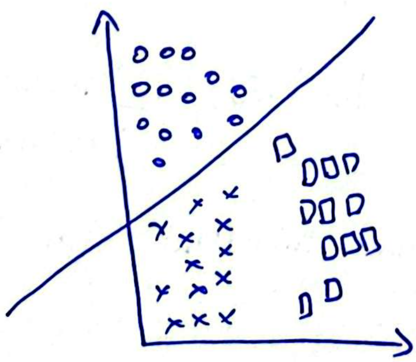
f_1	f_2	f_3	output	O_1	O_2	O_3
–	–	–	O_1	1	0	0
–	–	–	O_2	0	1	0
–	–	–	O_3	0	0	1
–	–	–	O_1	1	0	0
–	–	–	O_2	0	1	0
–	–	–	O_3	0	0	1



M1

Binary classification

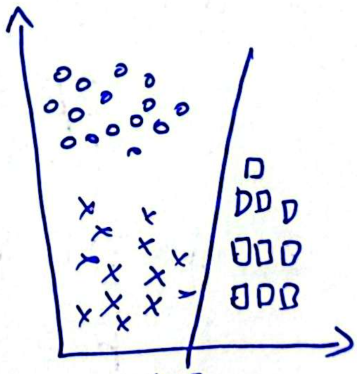
$f_1, f_2, f_3 \Rightarrow O_1$



M2

Binary classification

$f_1, f_2, f_3 \Rightarrow O_2$



M3

Binary classification

$f_1, f_2, f_3 \Rightarrow O_3$

New Test data → M1 → 0.25

→ M2 → 0.20

→ M3 → 0.55

$\Big\}$ [0.25, 0.20, 0.55]

→ final output (O_3)